

VU Research Portal

Thesaurus and Metadata Alignment for a Semantic E-Culture Application

Tordai, A.; Omelayenko, B.; Schreiber, A.T.

published in

Proc. 4th Int. conf. on Knowledge Capture, Whistler, Canada
2007

document version

Publisher's PDF, also known as Version of record

[Link to publication in VU Research Portal](#)

citation for published version (APA)

Tordai, A., Omelayenko, B., & Schreiber, A. T. (2007). Thesaurus and Metadata Alignment for a Semantic E-Culture Application. In *Proc. 4th Int. conf. on Knowledge Capture, Whistler, Canada* (pp. 199-200). ACM.
<http://www.cs.vu.nl/~guus/papers/Tordai07a.pdf>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:

vuresearchportal.ub@vu.nl

Thesaurus and Metadata Alignment for a Semantic E-Culture Application

Anna Tordai
VU University
1081a De Boelelaan
Amsterdam, Netherlands
atordai@cs.vu.nl

Borys Omelayenko
VU University
1081a De Boelelaan
Amsterdam, Netherlands
b.omelayenko@cs.vu.nl

Guus Schreiber
VU University
1081a De Boelelaan
Amsterdam, Netherlands
schreiber@cs.vu.nl

ABSTRACT

In this paper we describe a methodological approach for porting cultural repositories to the Semantic Web, focusing on the global picture of the required mappings and alignments.

Categories and Subject Descriptors: I.2.4 [Artificial Intelligence] Knowledge Representation Formalisms and Methods — *semantic networks*

General Terms: Standardization

Keywords: Semantic web, thesaurus alignment, schema mapping

1. INTRODUCTION AND APPROACH

This work is done in the context of the MultimediaN E-Culture project [1], the objective of which is to create a large virtual collection of cultural-heritage objects that supports semantic search. In this project we built a demonstrator¹ where multiple collections and vocabularies are converted to RDF/OWL and are aligned semantically. These include the leading Dutch art and ethnographic collections and vocabularies such AAT and TGN from Getty² and the Dutch ethnographic thesaurus SVCN³.

From these conversions we generalized a methodological approach to convert a new collection to RDF/OWL. Typically, a collection consists of two parts: the object descriptions (metadata) and an in-house vocabulary (thesaurus). We propose a four-step process summarized in Fig. 1 which includes the following steps:

¹<http://e-culture.multimedien.nl>

²<http://www.getty.edu/research/>

³<http://svcn.nl>

thesaurus conversion, metadata schema mapping, metadata mapping and thesaurus alignment.

In this paper we illustrate the process with examples of the Bibliopolis collection⁴ from the National Library of the Netherlands. The collection consists of 1,645 images related to book-printing accompanied by a thesaurus containing 1,033 terms used as keywords for indexing images. Both the thesaurus and the metadata are bilingual (English and Dutch).

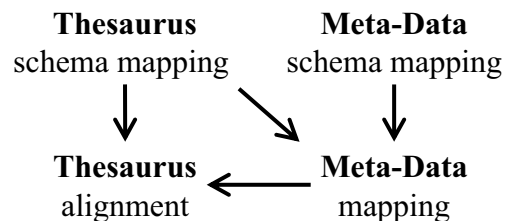


Figure 1: The four activities for converting a collection.

2. THESAURUS CONVERSION

Thesaurus schema mapping and conversion is a relatively well-researched area. In our work we use SKOS⁵ as the thesaurus schema and the method for thesaurus conversion proposed by van Assem [2].

3. METADATA SCHEMA MAPPING

In this activity we map the original record fields to the VRA Core scheme which is a specialization of Dublin Core⁶ for visual resources (our target type of resources).

Table 1 shows several conversion rules from the XML record fields to a VRA metadata schema with examples. Most of the target properties could be found in VRA, however, several specializations were needed such as `vra:subject.geographicPlace` which is a specialization of `vra:subject` as shown in Table 1 for field TWGEO.

⁴<http://www.bibliopolis.nl/>

⁵<http://www.w3.org/2004/02/skos/>

⁶<http://dublincore.org/>

| Field | Function | Conversion rule | Source value and target RDF/n3 |
|----------|---------------------------------|---|--|
| TITEL | Title in Dutch | Create literal and language tag | <i>source:</i> Delftse Bijbel... <i>target:</i> vra:title "Delftse Bijbel..."@nl ; |
| TITEL.EN | Title in English | Create literal and language tag | <i>source:</i> Delft Bible... <i>target:</i> vra:title "Delft Bible..."@en ; |
| MAKER | Creator and his marker for role | Extract name and role marker, create URI and label for name and convert marker to role, create role as subproperty of vra:creator | <i>source:</i> Yemantszoon, Mauricius : d <i>comment:</i> d stands for 'drukker' meaning 'printer' <i>target:</i> bp:drukker bp:Yemantszoon_Mauricius ; bp:Yemantszoon_Mauricius rdf:type ulan:person ; rdfs:label "Yemantszoon Mauricius" . |
| TWOND | Thesaurus term used as subject | Create mapping to thesaurus | <i>source:</i> typografische vormgeving <i>target:</i> vra:subject bp:typografische_vormgeving ; |
| TWGEO | Place used as subject for work | Create mapping to TGN where possible or keep literal | <i>source:</i> Delft <i>target:</i> vra:subject.geographicPlace tgn:7006804 ; |

Table 1: Part of the Bibliopolis metadata conversion rules

| Source Data | Vocabulary | Terms | | Instances | |
|------------------------|------------|--------|-------|-----------|-------|
| | | Mapped | Total | Mapped | Total |
| Thesaurus | AAT | 209 | 1033 | - | - |
| Metadata technique | AAT | 15 | 28 | 1332 | 1468 |
| Metadata object type | AAT | 14 | 19 | 978 | 1507 |
| Metadata subject place | TGN | 32 | 69 | 349 | 480 |

Table 2: Statistics for the Bibliopolis data and other vocabularies

4. METADATA VALUE MAPPING

After the schema is created the data values of the fields have to be converted. We have two kinds of fields: those that contain free-text literal values, such as the fields `vra:title` and `vra:description`, and those that contain values from (implicit) vocabularies, such as the fields for keywords or geographic places. In the latter case we distinguish between four kinds of vocabularies to which the field value can be mapped:

1. The local vocabulary such as an in-house thesaurus of keywords.
2. A standard external vocabulary such as AAT.
3. A vocabulary that is implicitly present in the field values, e.g. value 'I: S. Dali' of field `vra:creator`, where 'I' stands for his role as an illustrator and is part of a vocabulary of roles.
4. Terms that do belong to a vocabulary, which is either unknown or the alignment of the term to the vocabulary cannot be determined.

At this stage we either replace fields values with existing vocabulary terms (options 1 and 2), or we create new RDF resources to represent the terms (options 3 and 4) performing their alignment at the last stage.

5. THESAURUS ALIGNMENT

The local thesaurus and the newly created vocabularies extracted from the data need to be aligned with the

standard vocabularies. In the ontology mapping field, virtually all methods rely on the richness of relations between ontological concepts. In contrary, thesauri often use just broader/narrower or related relations. Therefore thesaurus alignment techniques need to rely heavily on label matching. For example, we aligned the Bibliopolis thesaurus to AAT by syntactically matching the Dutch `skos:prefLabel` to the Dutch translation of AAT preferred terms and mapped 209 concepts out of 1033 as presented in Table 2.

We use the SKOS Mapping Vocabulary specification⁷ created for the purpose of linking thesauri to each other with relationships `skos:exactMatch`, `skos:broadMatch`, etc. For this alignment the mappings are still based on the lexical match of term labels, that corresponds to the relation `skos:exactMatch`. Geographical names, however, form a frequent exception. With a few additional simple restrictions, a lexical match gives enough confidence to generate a semantic match as strong as `owl:sameAs`. As an example, a mapping to "Paris", known to be a city in France, can be made with `owl:sameAs`.

6. CONCLUSION

With thesauri being nearly the only way to link multiple repositories on the semantic web, we provide a practical methodology for aligning thesauri and metadata for the cultural heritage domain.

Acknowledgment. We are grateful to all our colleagues from the Multimedial E-Culture project, funded through the BSIK programme of the Dutch government.

7. REFERENCES

- [1] G. Schreiber et al. Multimedial e-culture demonstrator. In *International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 951–958. Springer, 2006.
- [2] M. van Assem, V. Malaisé, A. Miles, and G. Schreiber. A method to convert thesauri to SKOS. In volume 4011 of *Lecture Notes in Computer Science*, pages 95–109. Springer, 2006.

⁷<http://www.w3.org/2004/02/skos/mapping/spec/>